doi:10.11887/j.cn.202201023 http://journal.nudt.edu.cn

稀疏平衡变分自动编码器的文本特征提取*

车 蕾
(北京信息科技大学 信息管理学院,北京 100192)

摘 要:针对文本特征提取方面的高维数据特征区分度较低、基于规则的特征学习的自学习性能差、变分自动编码器存在过度剪枝等问题,提出稀疏平衡变分自动编码器(Sparse Balanced Variational AutoEncoder, SBVAE)的文本特征提取模型。为消除噪声干扰,提高文本特征提取模型的鲁棒性,在文本特征提取的输入层采用双向降噪处理机制。提出一种稀疏平衡性处理,结合 KL(Kullback-Leibler)项权重的模拟退火算法以缓解 KL 散度引发的过度剪枝的影响,强制解码器更充分地利用潜变量。此模型提高了高维数据特征的区分度。从对比分析文本特征提取模型、稀疏性能、稀疏平衡处理对隐藏空间变分下界的影响等方面深入开展实验,验证了该模型具有较好的性能。该模型在复旦数据集和 Reuters 数据集上的最高准确率相较于主成分分析分别提升了 12.36%、8.06%。

关键词:变分自动编码器;降噪;稀疏平衡;过度剪枝
中图分类号:TP391 **文献标志码**:A **文章编号**:1001-2486(2022)01-169-10

Text feature extraction based on sparse balanced variational autoencoder

CHE Lei
(School of Information Management, Beijing Information Science & Technology University, Beijing 100192, China)

Abstract: In order to solve the problems of low feature differentiation of high-dimensional data in text feature extraction, poor self-learning performance of rule-based representation learning, and excessive pruning of variational autoencoder, a text feature extraction model based on SBVAE (sparse balanced variational autoencoder) was proposed. In order to eliminate noise interference and improve robustness of the text feature extraction model, a bidirectional noise reduction mechanism was designed for variational autoencoder in the input layer of the text feature extraction. A sparse balance method combined with simulated annealing algorithm of weights of KL (Kullback-Leibler) terms was proposed to alleviate the effect of excessive pruning caused by KL divergence, and forced decoders to make full use of the latent variables. The model improves the discrimination of high-dimensional data features. Experiments were carried out in several aspects, including comparative analysis of text feature extraction model, sparse performance and influence of sparse balance on the lower bound of variation in hidden space. The results show that the proposed model has good performance. The highest accuracy of the proposed model of Fudan and Reuters datasets is increased by 12.36% and 8.06% in comparison with that of PCA, respectively.

Keywords: variational autoencoder; noise reduction; sparse balance; excessive pruning

随着各类语言模型层出不穷,人们开始将目光聚焦到模型最根本的表示层上。在机器学习中,特征学习(Feature Learning,FL)是将原始数据转换成机器学习能够处理的形式。原始数据通常为高维数据,高维数据既是维数福音,又是维数灾难,对机器学习算法提出挑战。特征学习的目的就是将高维的冗余的原始特征转换为低维的保留有效信息的特征[1]。特征学习包括特征选择和特征提取。区别于单纯选择子集的特征选择,特征提取是将原始表达投影到一个低维特征空间中以得到一个更加紧凑的表达[2]。本文研究工作是针对文本特征提取展开的。

目前研究存在如下问题:高维数据特征区分度较低、基于规则的特征学习的自学习和自适应能力差。深度学习则可以提高特征学习的自学习和自适应能力。本文将借助深度学习中变分自动编码器(Variational AutoEncoder,VAE)的潜变量研究文本特征提取的方法。潜变量产生即低维特

* **收稿日期**:2020-07-07
基金项目:北京市教育委员会社科计划一般项目(SM201911232003);北京信息科技大学教学改革项目重点资助项目(2020JGZD03);教育部人文社科规划基金资助项目(20YJAZH129)
作者简介:车蕾(1979—),女,河南洛阳人,副教授,博士,硕士生导师 E-mail:chelei@bistu.edu.cn

征表示,但是训练过程中存在潜在损失过度剪枝(即KL(Kullback-Leibler)散度为零)的情况。在这种情况下,生成器倾向于完全忽略潜在表示并简化为标准语言模型[3]。解决此问题的最典型方法是使用KL项权重的模拟退火算法来权衡重构误差与KL散度的贡献[4],但是此方法可能无法提高生成样本的质量,网络需要额外的容量来提高重构质量,代价是潜在空间利用率变低,很难被随机生成器利用。

为解决以上问题,本文提出一种稀疏平衡变分自动编码器(Sparse Balanced Variational AutoEncoder,SBVAE)的文本特征提取模型。本文简要说明变分自动编码器原理,分析降噪处理过程,详细阐述稀疏平衡性方法。本研究的实验选取代表性的真实数据集,从文本特征提取模型对比分析、稀疏性能分析、稀疏平衡处理对隐藏空间变分下界(Evidence Lower BOund,ELBO,也称为证据下界)的影响等几个方面深入开展,验证了SBVAE文本特征提取模型具有较好的性能。

1 相关工作

特征提取的常用传统算法有:概率模型、文档频率、信息增益等。传统方法特征识别度较低。基于概率模型的方法选择出来的词汇能够有效地代表类别,但可能会过滤掉代表性强但出现频率较低的词[5-6];基于文档频率的方法有可能丢弃某些有特殊含义的低频词条,筛选时需要选择恰当的阈值[7];基于信息增益的方法缺乏各特征词对特定文档的代表性考虑,实际应用效果很难达到理论效果[8-9]。

VAE模型是深度学习中的一种无监督表示学习和深度生成模型[10]。VAE具有与标准自动编码器(AutoEncoder,AE)完全不同的特性,它的隐含空间被设计为连续的分布以便进行随机采样和插值,这使得它成了有效的生成模型。VAE通过KL散度使潜变量分布靠近标准正态分布,从而能解耦潜变量特征,简化后续建立在该特征之上的模型,增强其泛化性能;另外,正态分布也使得编码空间更加规整。2013年,Kingma等提出VAE,它是深度生成模型,而不是基本的自动编码器[11]。

VAE模型在自然图像[12-13]和语音[14]生成方面取得了比较好的成果。目前,VAE模型在自然语言方面也取得一些进展。Wolf-Sonkin等基于VAE开展了上下文形态学拐点的研究[15]。Yang等基于VAE开展了文本建模的研究[16]。Li等基于VAE提出一种深度网络表示模型,无缝地集成文本信息和网络结构[17]。文献[16]和文献[18]基于VAE开展半监督分类的研究工作。Semeniuta等探讨了架构选择对学习文本生成的VAE的影响[3]。Louizos等基于VAE研究学习表示模型,该模型对于某些有害或数据变化的敏感因素是不变的,同时保留尽可能多的剩余信息[19]。目前VAE的研究主要是面向生成问题的,在文本特征提取方面的研究还比较少。VAE的主要优点是能够学习输入数据的平滑潜在状态表示,所以借助VAE的潜在空间研究文本特征提取问题是非常有价值的。但是,在足够高维度的潜在空间中,VAE存在过度剪枝的问题,倾向于忽略大量潜变量,即KL散度为零[20]。过度剪枝的做法可能会影响生成模型的质量。解决此问题的最典型方法是使用KL项权重的模拟退火算法来权衡重构误差与KL散度的贡献[4]。KL退火可以看成是从传统确定性自动编码器到完整VAE的逐渐过渡,但是这种方式很难被随机生成器利用。

2 模型

本文提出一种SBVAE文本特征提取模型。研究按如下步骤完成:

步骤1:采用词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)算法进行文本表示。

步骤2:针对数据量大、维度高的数据集,为提高数据的特征项明确度和分类精度,采用VAE进行文本特征提取。

步骤3:为提高鲁棒性,采取消除噪声干扰方法,在文本特征提取的输入层采用双向降噪处理机制。

步骤4:为缓解KL散度引发的过度剪枝的影响,并强制解码器更充分地利用潜变量,结合KL项权重的模拟退火算法提出一种稀疏平衡性处理方法。

步骤5:采用文本聚类算法验证文本特征提取的性能[21]。

2.1 变分自动编码器

自动编码器是利用人工神经网络的特点构造而成的网络[22]。自动编码网络,是由一组受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)按一定次序连接而构成的[22]。自动编码器的基本要求是保留原始数据的(尽可能多的)重要信息。当编码向量的分布尽可能接近高斯分布时,就导

致了 VAE 的产生。VAE 是一种将变分推理和深度学习相结合的潜变量生成模型,包括两部分:编码器和解码器[11]。相对于自动编码器,VAE 做了两大改进:第一,输入 $\boldsymbol{x}$ 的由编码器提供的确定性内部表示 $\boldsymbol{z}$ 被后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 替换。然后通过解码器将采样出来的 $\boldsymbol{z}$ 用于还原输入。为了简化采样过程,后验分布选择标准正态分布,其均值和方差由编码器预测。第二,为了确保可以从潜在空间的任何点进行采样并能产生有效、多样的输出,后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 通过其与分布 $p_\lambda(\boldsymbol{z})$ 的 KL 散度进行正则化。通常,先验分布也选择标准正态分布,使得可以以封闭形式计算先验和后验之间的 KL 散度。VAE 让所有的 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 接近标准正态分布,不仅防止了噪声为零的情况,也保证了模型具有生成能力。VAE 结构如图 1 所示。

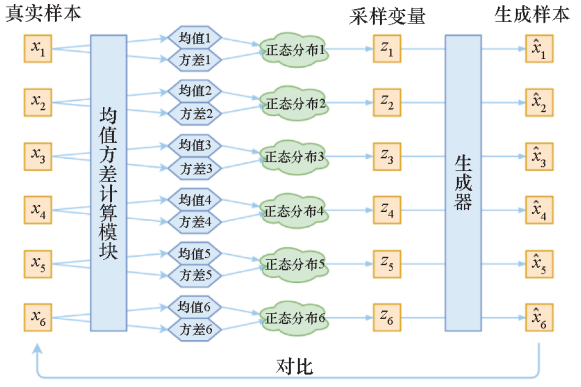

图 1　变分自动编码器

Fig. 1　Variational autoencoder

VAE 是概率潜变量模型,通过条件分布将观察到的向量 $\boldsymbol{x}$ 与低维潜变量 $\boldsymbol{z}$ 相关联[23]。VAE 模拟 $\boldsymbol{x}$ 的概率是:

$$p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\lambda(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \tag{1}$$

其中:$p_\theta(\boldsymbol{x}|\boldsymbol{z})$ 是给定 $\boldsymbol{z}$ 的 $\boldsymbol{x}$ 的条件分布,其由参数为 θ 的神经网络建模;$p_\lambda(\boldsymbol{z})$ 是潜变量的先验,其由参数为 λ 的神经网络建模。这些神经网络称为解码器。

对数似然估计 $\log p_\theta(\boldsymbol{x})$ 的下界是来自 Jensen 不等式的 ELBO[1],如式(2)所示。

$$\begin{aligned}\log p_\theta(\boldsymbol{x}) &= \log E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\lambda(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]\\ &\geqslant E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x}|\boldsymbol{z})p_\lambda(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right]\\ &\equiv \mathcal{L}(\boldsymbol{x};\theta,\phi)\end{aligned} \tag{2}$$

其中:$E[\,]$ 表示数学期望;后验分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 是在 $\boldsymbol{x}$ 发生的条件下 $\boldsymbol{z}$ 发生的概率,其由参数为 ϕ 的神经网络建模。$q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 通常为正态分布 $\mathrm{N}(\boldsymbol{z}|\mu_\phi(\boldsymbol{x}),\sigma_\phi^2(\boldsymbol{x}))$,$\mu_\phi(\boldsymbol{x})$ 和 $\sigma_\phi^2(\boldsymbol{x})$ 是参数为 ϕ、输入为 $\boldsymbol{x}$ 的神经网络。这些神经网络都称为编码器。此时 $\mathcal{L}(\boldsymbol{x};\theta,\phi)$ 可以看成 $\log p_\theta(\boldsymbol{x})$ 的下界。

ELBO 也可以写成如下形式:

$$\begin{aligned}\mathcal{L}_{\mathrm{VAE}} &= \mathcal{L}(\boldsymbol{x};\theta,\phi)\\ &= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,\|\,p_\lambda(\boldsymbol{z}))\end{aligned} \tag{3}$$

$$\begin{aligned}&D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,\|\,p_\lambda(\boldsymbol{z}))\\ &= \frac{1}{2}\sum_{k=1}^{D}(\sigma_{(k)}^2(\boldsymbol{x}) + \mu_{(k)}^2(\boldsymbol{x}) - \log\sigma_{(k)}^2(\boldsymbol{x}) - 1)\end{aligned} \tag{4}$$

其中,D 为 $\boldsymbol{x}$ 的样本数量。损失函数包括两项:第一项是惩罚重构误差 $E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$,由于要从分布里采样,所以相对自动编码器多了个期望值运算符;第二项是相对熵 $D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\,\|\,p_\lambda(\boldsymbol{z}))$,鼓励模型学习分布 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 近似于真实的先验分布 $p_\lambda(\boldsymbol{z})$,即一般正态分布[24]。重构的过程是希望没有噪声的,而 KL 散度则希望有高斯噪声,两者是对立的。KL 损失函数鼓励所有编码围绕隐藏层中心分布,同时惩罚不同分类被聚类到分离区域的行为。利用纯粹 KL 散度损失得到的编码是以隐藏空间为中心随机分布的。但是解码器从这些无意义的表达中很难解码出有意义的信息。KL 损失和重构损失结合起来就解决了这个问题。这使得在局域范围内的隐藏空间点维持了相同的类别,同时在全局范围内所有的点也被紧凑地压缩到了连续的隐含空间中。这一结果是通过重构损失的聚类行为和 KL 损失的紧密分布行为平衡得到的,从而形成了可供解码器解码的隐含空间分布。

2.2　降噪处理

为了提高 VAE 的鲁棒性和泛化性能,基于 VAE 从数据中学习到更有用的特征表示,本文对 VAE 进行双向降噪处理。在样本输入时加入了某种类型的噪声,使模型能够从受“污染”的部分输入中重构出“纯净”的输入,同时使模型能够从部分的“纯净”输入中反向重构出受“污染”的输入。即在降噪环节,首先随机选择一定比例的非 0 数据将其置为 0,即随机的损坏处理,从而加入噪声;然后随机选择更小比例的 0 将其置为 1,即随机的修复处理。这样可以使参数初始化尽量处于全局最优空间附近,避免进入局部最优参数空间。

2.3 稀疏平衡性处理

在足够高维的潜在空间中工作,网络学习表示的实际网络容量更加紧凑,许多潜变量独立于输入而被清零,并且完全被生成器忽略。当潜在空间的采样接近于1的时候,则认为它被激活,而采样接近于0的时候则认为它被抑制,使得神经元在大多数情况下都是被抑制的限制称为稀疏性限制。稀疏性限制可以让神经网络即使在隐藏神经元数量较多的情况下仍然可以发现输入数据的结构信息。VAE中的KL散度可以让高维潜在空间中的数据变得自然稀疏[25],它充当了正则化的作用,具备稀疏性相关的典型优势:它强制模型专注于真正重要的特征,大大降低过度拟合的风险。特别是,它是正确调整模型容量的主要方法,逐步训练模型以获得稀疏性或者减少网络的维度以移除未使用的神经元的连接。

研究发现,在足够高维度的潜在空间中,VAE存在过度剪枝的问题,倾向于忽略大量潜变量,即KL散度为零。过度剪枝[20]的做法可能会影响生成模型的质量。解决此问题最典型的方法是使用KL项权重的模拟退火算法来权衡重构误差与KL散度的贡献,如式(5)[26]所示,其中α表示超参数。

$$\mathcal{L}_{\mathrm{VAE}} = E_{q_\phi(z|x)}[-\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \alpha D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p_\lambda(\boldsymbol{z})) \tag{5}$$

KL退火可以看成是从传统确定性自动编码器到完整VAE的逐渐过渡,如式(5)所示,本研究也同时采取0到1的线性退火解决KL消失的问题。首先运行一个KL权重固定为0的模型,以找到它需要收敛的迭代次数;然后,将退火计划配置为在非正则化模型收敛后开始,持续时间不少于该数量的20%[3]。调低α可以控制不确定性,但是可能无法提高生成样本的质量,网络需要额外的容量来提高重构质量,代价是潜在空间利用率变低,很难被随机生成器利用。解决过度剪枝的另一个方法是修改模型架构,例如,文献[27]中提出了一种概率生成模型,该模型由许多称为缩影的稀疏变分自动编码器组成,这些缩影共享自动编码器的解码器架构。

本文采用一种平衡性方法缓解可能过度剪枝的影响。所提出的稀疏平衡是指为避免过度剪枝,在VAE的损失函数$\mathcal{L}_{\mathrm{VAE}}$中加入一个辅助稀疏惩罚项,平衡潜在空间变量的稀疏化。稀疏平衡使潜在空间的采样均值尽量接近0.01~0.03之间的某个值,以便达到稀疏的目的。整个过程肯定会丢失信息,但训练能够使丢失信息尽量少。添加了辅助稀疏惩罚项后的VAE损失函数表达式如式(6)所示。

$$\begin{aligned}\mathcal{L}_{\mathrm{SBVAE}} &= \mathcal{L}_{\mathrm{VAE}} + \beta\sum_{j=1}^{m} KL(\rho \| \bar{\rho}_j) \\ &= E_{q_\phi(z|x)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \alpha D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p_\lambda(\boldsymbol{z})) + \\ &\quad \beta\sum_{j=1}^{m} KL(\rho \| \bar{\rho}_j)\end{aligned} \tag{6}$$

其中:α和β表示超参数;m表示隐藏神经元数目;后一项为KL距离,其表达式如式(7)所示。

$$KL(\rho \| \bar{\rho}_j) = \rho\log\frac{\rho}{\bar{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\bar{\rho}_j} \tag{7}$$

其中:ρ是隐藏层节点的激活值,即目标稀疏值,通常是一个接近于0的较小的数;$\bar{\rho}_j$项(如式(8)所示)是隐藏层中隐藏神经元j在整个训练集上的平均激活度,在梯度下降优化函数上增加限制因素。

$$\bar{\rho}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ji} \tag{8}$$

其中:n表示输入向量数;z_{ji}表示隐藏神经元j对应第i个向量的激活度。为避免$\boldsymbol{z}$为负值引起的$\mathcal{L}_{\mathrm{SBVAE}}$无穷大,在输出之前需要对$\boldsymbol{z}$值取sigmoid后再进行稀疏平衡处理。$\mathcal{L}_{\mathrm{SBVAE}}$的目标函数对潜变量$\boldsymbol{z}$施加了约束,以产生有效特征。

3 实验

实验分别从实验环境及数据、对比实验及参数设置、性能评价指标、实验结果与分析四个方面进行详细讨论。

3.1 实验环境及数据

为了验证模型的有效性,采用如下2种标准数据集(均可以通过开源网站获得)进行实验。

1)复旦数据(中文)[28]:复旦大学文本分类数据,此数据集由复旦大学计算机信息与技术系国际数据库中心自然语言处理小组提供。实验从中选取了3个话题,记录数为8 423条。实验数据集的分布情况如图2所示。

2)路透社(英文)[29]:此新闻数据集是由路透公司采集的1987年的新闻稿组成的Reuters-21578文集作为实验数据集。实验从中选取了3个话题,记录数为6 740条。实验数据集的分布情况如图3所示。

对采集到的中文文本进行如下预处理:首先采用结巴(jieba)分词工具对数据进行分词,并去除文本中的数字;接着过滤掉叹词、副词等相关性较弱的词语和标点符号;最后去除停用词。对采

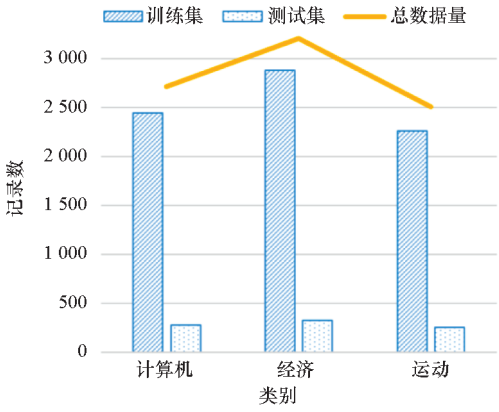

图 2　复旦数据集数据分布

Fig. 2　Data distribution of Fudan dataset

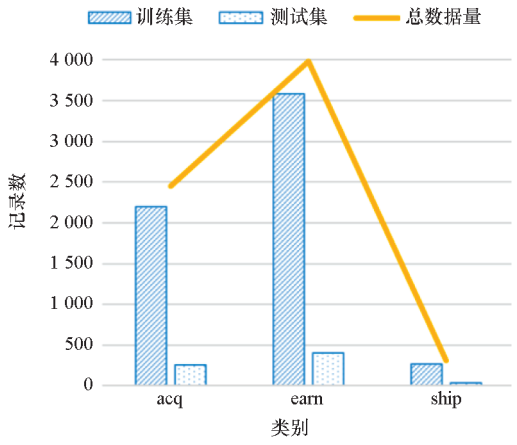

图 3　路透社数据集数据分布

Fig. 3　Data distribution of Reuters dataset

集到的英文文本进行如下预处理:去除数字和停用词;字母全部转为小写;用 Porter 算法进行词干化处理,将英文中单复数、时态等变形词转换成原型。所有字符编码采用无 BOM 的 UTF－8 格式。

3.2　对比实验及参数设置

本文选取了 6 个模型进行对比实验,以便验证本文提出的模型的性能。具体模型信息如下:①主成分分析(Principal Component Analysis,PCA);②AE;③稀疏自动编码器(Sparse AutoEncoder,SAE);④VAE;⑤降噪变分自动编码器(Denoising Variational AutoEncoder,DVAE);⑥SBVAE。

本实验的训练参数是通过多次实验获得的最优参数。因为实验涉及的是多分类问题,所以交叉熵采用的是 Categorical_Crossentropy。最大 epoch 为 500,基础学习率为 0.001。实验训练采用的是自收敛方式,当损失值变化连续 n 次都低于某个临界值(默认 0.000 1)时,训练自动终止。编码器和解码器都采用完全连接网络。模型采用 Keras 的自适应优化器 Adadelta,允许在每个步骤中为学习率计算不同的值,以使训练效果达到最优。

3.3　性能评价指标

实验通过采用 K-means 聚类算法验证特征提取的性能。性能评价指标包括:准确率、召回率、F1 值、熵和纯度。准确率、召回率、F1 值和纯度越高越好,熵值越低越好。

1)准确率:预测正确的结果占总样本的百分比。

2)召回率:在实际为某类的样本中被预测为该类样本的概率。

3)F1 值:同时考虑准确率和召回率,让两者同时达到最高,取得平衡。计算公式如下:

$$\text{F1 值} = 2 \times \frac{\text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \tag{9}$$

4)熵:对于一个聚类 i,首先计算 p_{ij}。$p_{ij} = \frac{m_{ij}}{m_i}$ 指的是聚类 i 中的成员属于类 j 的概率,其中 m_i 是聚类 i 中所有成员的个数,m_{ij} 是聚类 i 中的成员属于类 j 的个数。每个聚类的熵如式(10)所示,其中 L 是类的个数。整个聚类划分的熵如式(11)所示,其中 K 是聚类的数目,m 是整个聚类划分所涉及的成员个数。

$$e_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij} \tag{10}$$

$$e = \sum_{i=1}^{K} \frac{m_i}{m} e_i \tag{11}$$

5)纯度:使用上述熵中的 p_{ij} 定义,聚类 i 的纯度定义如式(12)所示,整个聚类划分的纯度如式(13)所示。

$$p_i = \max(p_{ij}) \tag{12}$$

$$p = \sum_{i=1}^{K} \frac{m_i}{m} p_i \tag{13}$$

3.4　实验结果与分析

通过比较实验和可视化实验,将已有的算法与本文提出的 SBVAE 算法进行对比,展示本算法的优势;通过实验,展示基于 SBVAE 的文本特征表示对聚类过程的贡献,以更好支持后续话题检测与追踪工作的研究。

3.4.1　实验结果对比分析

本节是基于 SBVAE 文本特征提取模型开展的聚类实验。表 1 是基于复旦数据集的实验结果比对,表 2 是基于路透社数据集的实验结果比对。从表 1 可以看出,SBVAE 的 F1 值依次比 PCA、AE、VAE 的 F1 值提升了 12.36%、1.12%、0.47%。从表 2 可以看出,SBVAE 的 F1 值依次比 PCA、AE、VAE 的 F1 值提升了 8.06%、

5.07%、3.42%。实验结果显示,不论是复旦数据集,还是路透社数据集,本文提出的 SBVAE 文本特征提取方法都能得到纯度较高的聚类效果,证实了该方法具有一定的稳定性。

表 1 复旦数据的实验结果对比

Tab. 1 Comparison of experimental results of Fudan dataset

模型	准确率/%	召回率/%	F1 值	熵	纯度
PCA	0.894 9	0.820 8	0.856 3	0.575 1	0.838 2
AE	0.950 2	0.952 6	0.951 4	0.308 3	0.951 0
VAE	0.957 7	0.957 6	0.957 6	0.274 8	0.958 6
DVAE	0.958 5	0.958 3	0.958 4	0.267 9	0.959 4
SBVAE	0.961 2	0.963 0	0.962 1	0.254 0	0.962 1

表 2 路透社数据的实验结果对比

Tab. 2 Comparison of experimental results of Reuters dataset

模型	准确率/%	召回率/%	F1 值	熵	纯度
PCA	0.875 4	0.577 5	0.695 9	0.697 2	0.791 1
AE	0.890 1	0.598 4	0.715 7	0.626 2	0.829 1
VAE	0.898 9	0.610 4	0.727 1	0.586 0	0.851 0
DVAE	0.888 3	0.625 1	0.733 8	0.461 8	0.898 2
SBVAE	0.916 7	0.637 5	0.752 0	0.474 2	0.902 8

图 4 和图 5 基于复旦数据集分别展示了二维空间下原始标注数据的分布和 SBVAE 特征提取(F1 值为 0.962 1)后 *K*-means 聚类效果。可以发现,图 5 的聚类效果较好,纯度较高,说明本文提出的 SBVAE 特征提取模型具有较好的性能。

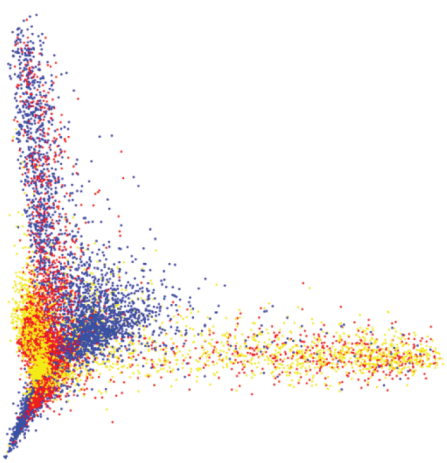

图 4 原始标注的复旦数据

Fig. 4 Fudan dataset of original annotation

3.4.2 稀疏性实验

为分析本文提出的模型的稀疏处理效果,将

图 5 SBVAE 特征提取后 *K*-means 聚类效果图

Fig. 5 *K*-means Clustering effect after feature extraction based on SBVAE

基于稀疏的自动编码器和基于稀疏的变分自动编码器进行对比分析。图 6 ~ 9 展示的是 SAE 中 $\bar{\rho}_j$ 的训练轨迹($\rho=0.01$ 和 $\rho=0.05$)、SBVAE 中 $\bar{\rho}_j$ 的训练轨迹($\rho=0.01$ 和 $\rho=0.05$)。由于 SBVAE 中是由 2 个 KL 散度(详见式(6))共同作用,训练过程中一方面向高斯分布逼近,一方面向 ρ 逼近,因此在训练过程中 $\bar{\rho}_j$ 向 ρ 的逼近程度没有 SAE 明显,上下对抗波动比较大。

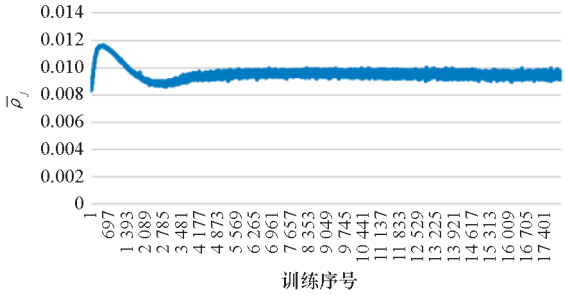

图 6 SAE 中 $\bar{\rho}_j$ 的训练轨迹($\rho=0.01$)

Fig. 6 Training track of $\bar{\rho}_j$ in SAE($\rho=0.01$)

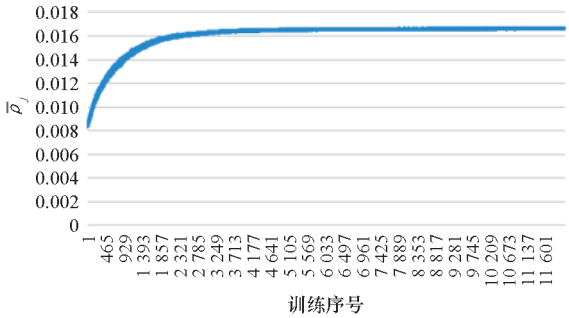

图 7 SAE 中 $\bar{\rho}_j$ 的训练轨迹($\rho=0.05$)

Fig. 7 Training track of $\bar{\rho}_j$ in SAE($\rho=0.05$)

图 10 ~ 19 显示的实验结果取的是 VAE 和 SBVAE 训练结束时的各文本隐藏空间对应的潜

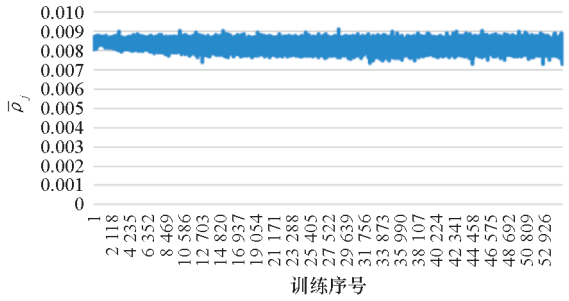

图 8　SBVAE 中 $\bar{\rho}_j$ 的训练轨迹($\rho=0.01$)

Fig. 8　Training track of $\bar{\rho}_j$ in SBVAE($\rho=0.01$)

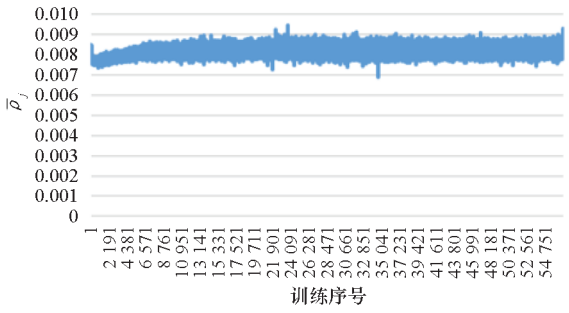

图 9　SBVAE 中 $\bar{\rho}_j$ 的训练轨迹($\rho=0.05$)

Fig. 9　Training track of $\bar{\rho}_j$ in SBVAE ($\rho=0.05$)

变量 z、方差和均值。潜变量 z、方差和均值都为向量,因此这里取的是向量各维度的均值构建的图例。相关公式如下:

$$\begin{cases} z\_mean_{(k)} = mean(z_{(k)}) \\ \sigma\_mean_{(k)} = mean(\sigma_{(k)}) \\ \mu\_mean_{(k)} = mean(\mu_{(k)}) \end{cases} \tag{14}$$

从图 10、图 11、图 15 和图 16 大致可以看出,稀疏平衡之前的潜变量 z 在 0 附近的密集程度要略低于稀疏平衡之后的潜变量 z 在 0 附近的密集程度。本实验的潜变量 z 的个数为 7 524,隐藏空间维度为 50,总向量值个数为376 200。实验统计,稀疏平衡之前潜变量 z 的向量值在 -1 与 1 之间的个数为 91 601 个,稀疏平衡之后潜变量 z 的向量值在 -1 与 1 之间的个数为93 830个。由此说明,稀疏平衡后潜变量 z 的稀疏程度大于稀疏平衡之前的稀疏程度。其中,排列图是指按频数的降序绘制数据的分布,累积线位于次坐标轴上,标识占总数的百分比。排列图的箱宽度是通过使用 Scott 正态引用规则计算的。从图 12、图 13、图 17 和图 18 稀疏平衡处理前后 VAE 中的方差分布情况可以看出,稀疏平衡后的潜在空间的方差逼近 1 的值也相对增多。从图 14 和图 19 稀疏平衡处理前后 VAE 中的均值分布情况可以看出,稀疏平衡后的均值逼近 0 的值也相对增多。

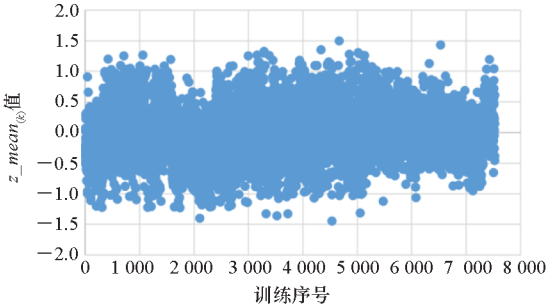

图 10　VAE 的潜变量 $z\_mean_{(k)}$ 分布(散点图)

Fig. 10　Distribution of latent variables $z\_mean_{(k)}$ in VAE (scatter plot)

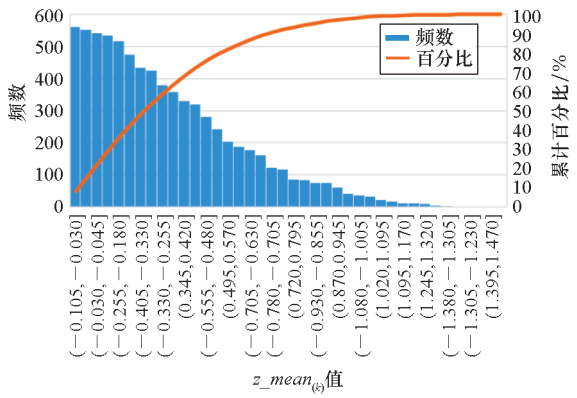

图 11　VAE 的潜变量 $z\_mean_{(k)}$ 分布(排列图)

Fig. 11　Distribution of latent variables $z\_mean_{(k)}$ in VAE (pareto diagram)

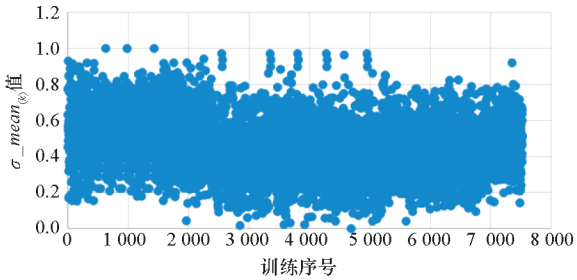

图 12　VAE 的 $\sigma\_mean_{(k)}$ 分布(散点图)

Fig. 12　Distribution of $\sigma\_mean_{(k)}$ in VAE(scatter plot)

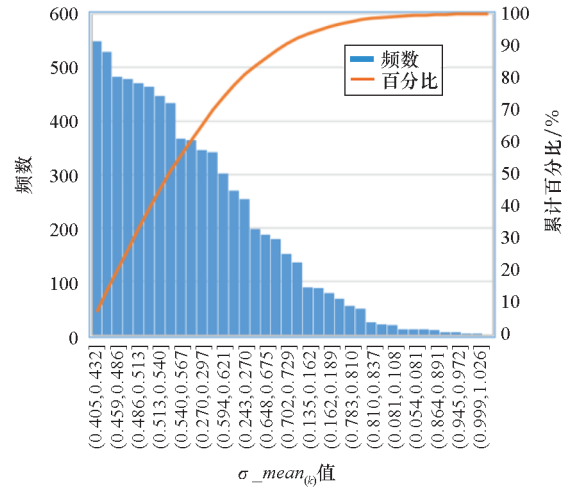

图 13　VAE 的 $\sigma\_mean_{(k)}$ 分布(排列图)

Fig. 13　Distribution of $\sigma\_mean_{(k)}$ in VAE (pareto diagram)

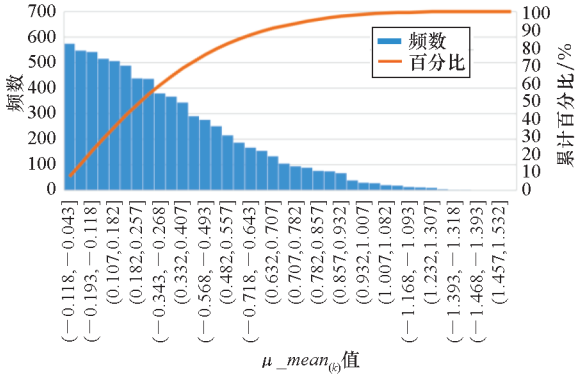

图 14　VAE 的 $\mu\_mean_{(k)}$ 分布(排列图)

Fig. 14　Distribution of $\mu\_mean_{(k)}$ in VAE (pareto diagram)

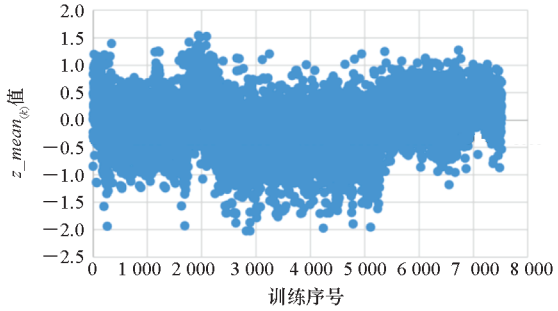

图 15　SBVAE 的潜变量 $z\_mean_{(k)}$ 分布(散点图)

Fig. 15　Distribution of latent variables $z\_mean_{(k)}$ in SBVAE(scatter plot)

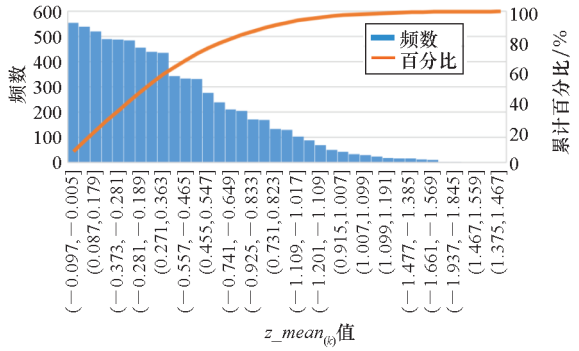

图 16　SBVAE 的潜变量 $z\_mean_{(k)}$ 分布(排列图)

Fig. 16　Distribution of latent variables $z\_mean_{(k)}$ in SBVAE (pareto diagram)

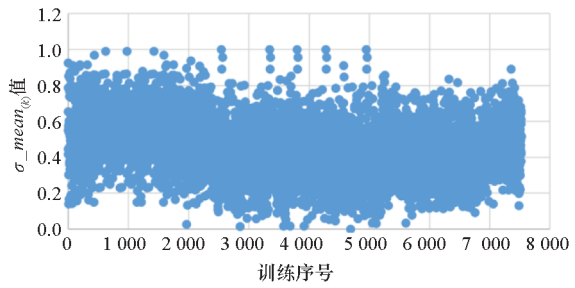

图 17　SBVAE 的 $\sigma\_mean_{(k)}$ 分布(散点图)

Fig. 17　Distribution of $\sigma\_mean_{(k)}$ in SBVAE (scatter plot)

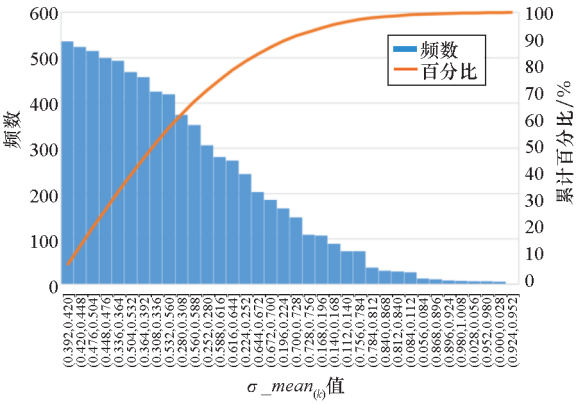

图 18　SBVAE 的 $\sigma\_mean_{(k)}$ 分布(排列图)

Fig. 18　Distribution of $\sigma\_mean_{(k)}$ in SBVAE (pareto diagram)

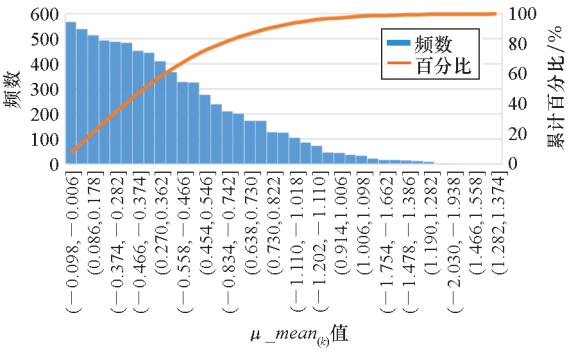

图 19　SBVAE 的 $\mu\_mean_{(k)}$ 分布(排列图)

Fig. 19　Distribution of $\mu\_mean_{(k)}$ in SBVAE (pareto diagram)

综上所述,稀疏平衡后的特征提取性能优于稀疏平衡之前的性能。

3.4.3　稀疏平衡处理对隐藏空间变分下界的影响

为了验证模型引入稀疏平衡处理对变分下界的影响,表 3 显示了基于复旦数据集在 SBVAE 中设置隐藏空间维度为不同值时变分下界的变化情况。其中,惩罚重构误差项、KL 散度 1 和 KL 散度 2 分别对应式(6)中等号右边的三个项,KL 散度 2 即辅助稀疏惩罚项。

表 3　SBVAE 的变分下界

Tab. 3　The ELBO of SBVAE

隐藏空间	$-\log p_\theta(x) \leqslant$	惩罚重构误差项	KL 散度 1	KL 散度 2
10 维	38.025 0	36.683 4	1.335 8	0.005 9
20 维	40.390 8	38.992 4	1.391 5	0.006 9
30 维	38.823 9	37.560 6	1.254 3	0.009 0
40 维	42.403 4	41.006 9	1.389 1	0.094 9
50 维	37.950 5	36.605 4	1.337 2	0.007 9
60 维	43.459 8	42.009 4	1.436 9	0.013 4

表 4 显示了基于复旦数据集在 VAE 中设置隐藏空间维度为不同值时变分下界的变化情况。其中,惩罚重构误差项和 KL 散度分别对应式(5)中等号右边的两个项。变分下界是越低越好。和 VAE 的变分下界相比,引入稀疏平衡处理后的 VAE 对变分下界($-\log p_\theta(x) \leqslant$)有显著的提升,其变分下界优于标准的 VAE 模型的变分下界。

表 4　VAE 的变分下界

Tab. 4　The ELBO of VAE

隐藏空间	$-\log p_\theta(x) \leqslant$	惩罚重构误差项	KL 散度
10 维	55.680 6	54.422 7	1.257 8
20 维	67.237 4	65.684 4	1.553 0
30 维	56.365 6	55.145 4	1.220 2
40 维	49.912 4	48.624 4	1.288 0
50 维	58.176 5	56.615 4	1.561 1
60 维	73.079 3	71.204 2	1.875 1

4　结论

本文提出了 SBVAE 文本特征提取模型。为提升高维数据的区分度,基于 VAE 研究文本特征提取问题;为消除噪声干扰,在输入层采用双向降噪处理机制;为缓解过度剪枝的影响,结合 KL 项权重的模拟退火算法提出稀疏平衡性处理,强制解码器更充分地利用潜变量。实验从多个方面深入开展,验证了 SBVAE 模型在文本特征提取问题的解决上具有较好的性能,对文本特征提取问题的研究具有一定的推动作用。

未来的工作方向将尝试把模型扩展到半监督学习中,学习潜变量的深度结构性分层,并进一步研究其推理方案。

参考文献(References)

[1] ABID A, BALIN M F, ZOU J. Concrete autoencoders for differentiable feature selection and reconstruction[EB/OL]. (2019-01-27)[2020-06-01]. https://arxiv.org/abs/1901.09346

[2] BANDHAKAVI A, WIRATUNGA N, PADMANABHAN D, et al. Lexicon based feature extraction for emotion text classification[J]. Pattern Recognition Letters, 2017, 93: 133-142.

[3] SEMENIUTA S, SEVERYN A, BARTH E. A hybrid convolutional variational autoencoder for text generation[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2017: 627-637.

[4] REZENDE D J, MOHAMED S. Variational inference with normalizing flows[C]// Proceedings of the 32nd International Conference on Machine Learning, 2015: 1530-1538.

[5] UYSAL A K, GUNAL S. The impact of preprocessing on text classification[J]. Information Processing & Management, 2014, 50(1): 104-112.

[6] LEE C H, WU C H, CHIEN T F. BursT: a dynamic term weighting scheme for mining microblogging messages[C]// Proceedings of 8th International Symposium on Neural Network, 2011: 548-557.

[7] SHI K S, HE J, LIU H T, et al. Efficient text classification method based on improved term reduction and term weighting[J]. Journal of China Universities of Posts and Telecommunications, 2011, 18(Suppl 1): 131-135.

[8] SHANG C X, LI M, FENG S Z, et al. Feature selection via maximizing global information gain for text classification[J]. Knowledge-Based Systems, 2013, 54: 298-309.

[9] JAVED K, MARUF S, BABRI H A. A two-stage Markov blanket based feature selection algorithm for text classification[J]. Neurocomputing, 2015, 157: 91-104.

[10] 杨任农, 房育寰, 张振兴, 等. 变分自编码器结合聚类算法在空战态势评估问题上的应用[J]. 国防科技大学学报, 2019, 41(4): 144-155.
YANG R N, FANG Y H, ZHANG Z X, et al. Application of variational autoencoder combined with clustering algorithm in air combat situation assessment[J]. Journal of National University of Defense Technology, 2019, 41(4): 144-155. (in Chinese)

[11] KINGMA D P, WELLING M. Auto-encoding variational Bayes[EB/OL]. (2013-12-20)[2020-06-01]. https://arxiv.org/abs/1312.6114.

[12] Bachman P. An architecture for deep, hierarchical generative models[C]// Proceedings of 30th Conference on Neural Information Processing Systems, 2016: 4833-4841.

[13] GULRAJANI I, KUMAR K, AHMED F, et al. PixelVAE: a latent variable model for natural images[C]// Proceedings of 5th International Conference on Learning Representations, 2017: 1-9.

[14] FRACCARO M, SØNDERBY S K, PAQUET U, et al. Sequential neural models with stochastic layers[C]// Proceedings of 30th Conference on Neural Information Processing Systems, 2016: 1-9.

[15] WOLF-SONKIN L, NARADOWSKY J, MIELKE S J, et al. A structured variational autoencoder for contextual morphological inflection[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 2631-2641.

[16] YANG Z C, HU Z T, SALAKHUTDINOV R, et al. Improved variational autoencoders for text modeling using dilated convolutions[C]// Proceedings of the 34th International Conference on Machine Learning, 2017: 3881-3890.

[17] LI H, WANG H Z, YANG Z L, et al. Variation autoencoder based network representation learning for classification[C]// Proceedings of ACL, Student Research Workshop, 2017: 56-61.

[18] XU W D, SUN H Z, DENG C, et al. Variational autoencoders for semi-supervised text classification[EB/OL]. (2016 - 03 - 08)[2020 - 06 - 01]. https://arxiv.org/abs/1603.02514.

[19] LOUIZOS C, SWERSKY K, LI Y J, et al. The variational fair autoencoder[EB/OL]. (2015 - 09 - 03)[2020 - 06 - 01]. https://arxiv.org/abs/1511.00830v6.

[20] ASPERTI A. Sparsity in variational autoencoders[EB/OL]. (2018 - 12 - 18)[2020 - 06 - 01]. https://arxiv.org/abs/1812.07238.

[21] 车蕾, 杨小平. 多特征融合文本聚类的新闻话题发现模型[J]. 国防科技大学学报, 2017, 39(3): 85 - 90.
CHE L, YANG X P. News topic discovery model of multi feature fusion text clustering [J]. Journal of National University of Defense Technology, 2017, 39(3): 85 - 90. (in Chinese)

[22] WANG S, DING Z, FU Y. Feature selection guided auto-encoder[C]// Proceedings of Thirty-First AAAI Conference on Artificial Intelligence, 2017: 2725 - 2731.

[23] TAKAHASHI H, IWATA T, YAMANAKA Y, et al. Variational autoencoder with implicit optimal priors [C]// Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019, 33(1): 5066 - 5073.

[24] LI P J, WANG Z H, LAM W, et al. Salience estimation via variational auto-encoders for multi-document summarization[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017: 3497 - 3503.

[25] YEUNG S, KANNAN A, DAUPHIN Y, et al. Tackling over-pruning in variational autoencoders[EB/OL]. (2017 - 06 - 09)[2020 - 06 - 01]. https://arxiv.org/abs/1706.03643.

[26] MESCHEDER L, NOWOZIN S, GEIGER A. Adversarial variational Bayes: unifying variational autoencoders and generative adversarial networks[C]// Proceedings of the 34th International Conference on Machine Learning, 2017: 2391 - 2400.

[27] YEUNG S, KANNAN A, DAUPHIN Y, et al. Epitomic variational autoencoder[C]// Proceedings of 5th International Conference on Learning Representations, 2017: 1 - 16.

[28] SCHWENK H. Continuous space language models [J]. Computer Speech & Language, 2007, 21(3): 492 - 518.

[29] PAUL S, MAGDON-ISMAIL M, DRINEAS P. Feature selection for linear SVM with provable guarantees[J]. Pattern Recognition, 2016, 60: 205 - 214.